ArkN4 og XML-parsing

26. mai 2011

Hallstein Bakken

Seksjon for digitalt depot

Riksarkivet , Oslo

ARKIVVERKET
RIKSARKIVET

Norsk lovverk

- **Overordnet premissgiver ved utvikling av testverktøy er norsk lovverk**
  - Utgangspunktet er Arkivloven
  - Til Arkivloven er det utarbeidet en forskrift
  - Innenfor vårt område gjelder kapittel VIII og IX
  - Kapittel VIII: Bestemmelser om elektronisk arkivmateriale som avleveres eller overføres som depositum til Arkivverket
  - Kapittel IX: Elektronisk arkivering av saksdokumenter

Metode

- **Spesifikke premissgivere er Noark-standardene**
  - Noark-3 : Samlet kravspesifikasjon (1994)
  - Noark-4 : Del 1 – Funksjonsrettet beskrivelse og kravspesifikasjon (1999)
    Del 2 – Tekniske spesifikasjoner (1999)
  - Noark-4.1 : Revidert kap. 15 (2002)
  - Noark 5 : Versjon 1.0 (4.7.2008)
    Versjon 1.2 (17.10.2008)
    Versjon 2.0 (3.4.2009)
    Versjon 2.1 (3.5.2010)
    Versjon 3.0 (1.3.2011)
- Skjemaer med testpunkter (Noark-3/Noark 5)
- For Noark-4 finnes kravspesifikasjon bestående av 22 punkter
- Metode for beskrivelse av fagsystemer: ADDML

Om standarden Noark-4

- **Komplisert struktur**
  - Antall avleverte tabeller vil variere
  - 95 mulige tabeller med til sammen 885 felter
  - 29 tabeller er obligatoriske iht. O-kravet
    - 215 felter er her obligatoriske
  - 39 tabeller obligatoriske iht. O2-kravet
  - Typisk avlevering inneholder 40-50 tabeller
- **Standardisert avleveringsformat**
  - Opptil 95 tabeller avleveres som XML-filer
  - Detaljert spesifisert med DTDer
  - Alle elektroniske saksdokumenter tilknyttet journalpostene skal følge med
  - Rapporter (også XML) skal også avleveres
  - Godkjenningstest av prøve på Noark-4 avleveringsuttrekk

Verktøyutvikling og testing

- **Utvikling av våre egne verktøy for testing av arkivversjoner.**
- **Noark 3 testes med Proteus.**
- **Noark 4 testes med ArkN4.**
- **Noark 5 testes med Arkade.**
- **Fagsystem (annet enn Noark) testes med Arkade.**

ArkN4

- **Hva er ArkN4?**

  - Et verktøy for å teste og tilgjengeliggjøre arkivuttrekk fra Noark-4

Om ArkN4 (1)

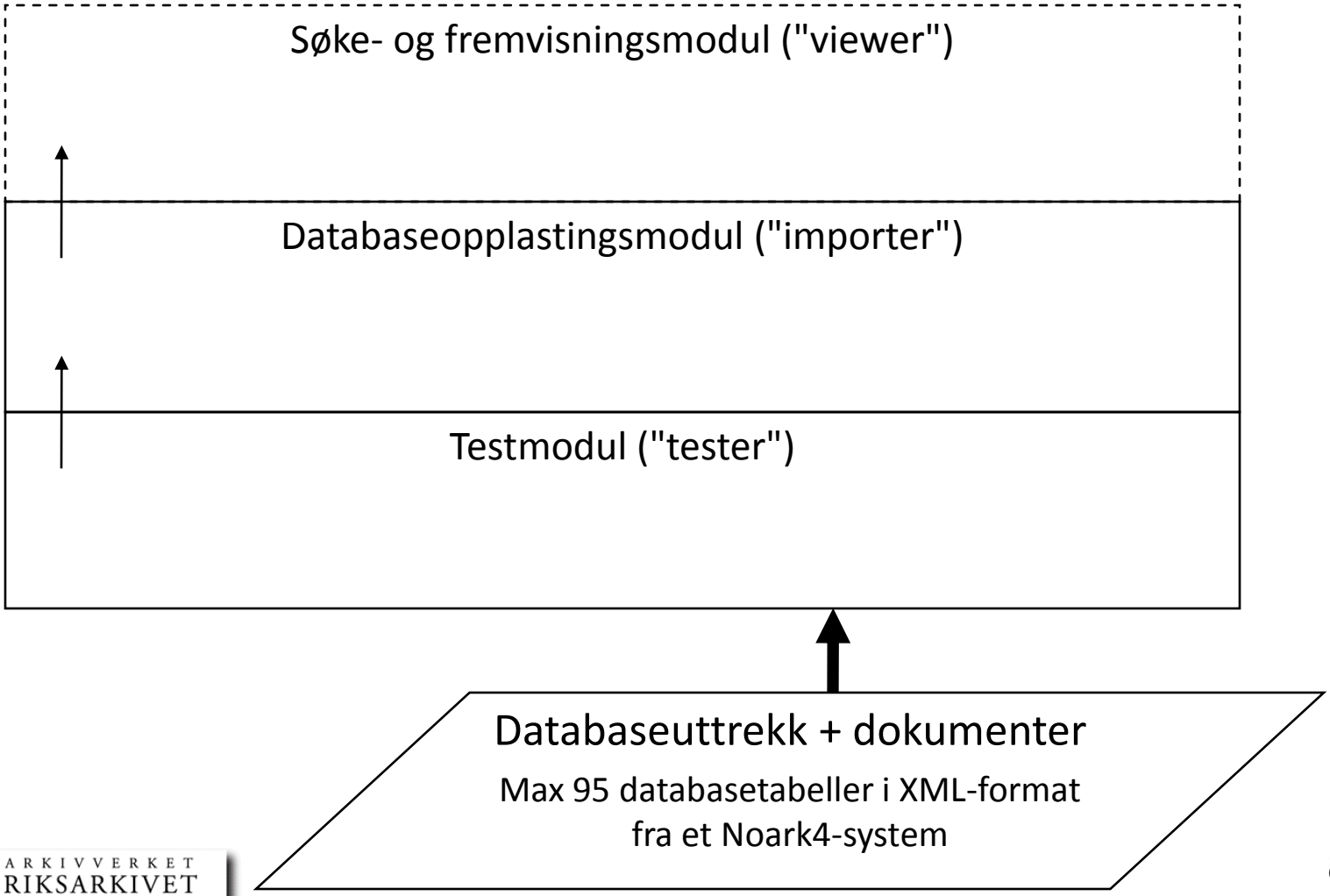
ArkN4 består av tre moduler

- XML-filene i databaseuttrekket leses en etter en inn i *testmodulen*, der de testes mot DTD'en for den aktuelle filen. Andre tester utføres også.
- *Databaseopplastingsmodulen* oppretter først en tom SQL-database basert på datamodellen for Noark-4, og importerer deretter dataene fra de enkelte XML-filene inn i databasen.
- *Søke- og fremvisningsmodulen* skal tilby ulike funksjoner for navigering og søking i den opprettede databasen. Skal gjøre det mulig å vise innholdet på en måte som ligner visnings-måten i det originale systemet.

  NB! Denne modulen er så langt ikke realisert.

Om ArkN4 (2)

Testing og tilgjengeliggjøring av uttrekk fra Noark-4



ARKIVVERKET
RIKSARKIVET

Om ArkN4 (3)

- ArkN4 er Åpen Programvare (Open Source)
  - Programmert i Perl på Linux, men skal også kunne kjøre på UNIX, Windows og Mac
  - Bruker databasen MySQL, men skal lett kunne tilpasses til andre relasjonsdatabaser
  - *Søke- og fremvisningsmodulen* skal utvikles i PHP og vil være web-basert dvs. at det kan kjøres i en vanlig nettleser uten installasjon av annen programvare på brukerens PC.

Om ArkN4 (4)

- Om programkoden: kompleks!
- ArkN4 består av:
  - 33 moduler
  - 302 funksjoner
  - 11973 linjer perlkode
  - 4555 linjer SQL, for å definere databasen
  - tilsammen 16528 linjer kode.
- XML-parsere
  - *tree* based parser
  - *stream* based parser
  - XML::Twig

”Tolerant” import til ArkN4

- Et generelt testverktøy for Noark-4-uttrekk må være tolerant overfor avvik og mangler i uttrekkene!
- Utfordrende å lage et robust testverktøy
- Vanskelig å lage et program som godtar ”alt” uten å stoppe.

ArkN4's virkemåte

- All parsing og testing av XML-filer skjer først. Deretter kan databaseimport utføres om ønskelig.
- Referanseintegritet kan skrues av under import til database.
- Mange "små-feil" godtas. Eks. case sensitivitet
- Umulig å se for seg alle mulige feil som avleverte XML-filer kan inneholde. Derfor: mye prøving og feiling under programutviklingen.

Kravspesifikasjon (1)

- **Kravspesifikasjonen for ArkN4**
  - Funksjonaliteten i ArkN4, Riksarkivets testverktøy for Noark 4-uttrekk, er utviklet på grunnlag av en kravspesifikasjon som ble utarbeidet av Jon Atle Haugen i daværende *Avdeling for elektronisk arkiv* i Riksarkivet i 2005.

  - Kravspesifikasjonen inneholder 22 punkter.

Kravspesifikasjon (2)

1. Validering av XML-filene mot de tilhørende DTDene, som følger med ArkN4

Dersom det oppdages feil, skal det skrives ut feilmeldinger.

Det hender at det følger egne DTDer med et uttrekk, men det er viktig å presisere at ArkN4 alltid skal validere XML-filene mot de offisielle DTDene. Unntaket er hvis det i uttrekket følger med egne filer med virksomhetsspesifikk informasjon. Da må det også følge med egne DTDer til disse filene.

2. Kontroll av at NOARKIH-filen (NOARKIH.XML) er med i uttrekket

Feilmelding skal skrives ut dersom filen mangler.

Kravspesifikasjon (3)

3. Kontroll av at de to rapportfilene SAKDOK og JOURNAL er med i uttrekket

Feilmelding skal skrives ut dersom filene mangler.

4. Kontroll av hvilke tabellfiler som er med i uttrekket, og om disse er spesifisert i NOARKIH-filen

I Noark 4-standarden er det spesifisert til sammen 95 tabeller. 29 tabeller inneholder obligatoriske dataelementer ved journalføring uten elektroniske dokumenter (O-krav). 39 tabeller inneholder obligatoriske dataelementer ved journalføring med elektroniske dokumenter (O- og O2-krav).

Det skal skrives ut feilmelding hvis det følger med filer som ikke er spesifisert i NOARKIH.

Kravspesifikasjon (4)

5. Kontroll av hvilke tabellfiler som *ikke* er med i uttrekket, men som er spesifisert i NOARKIH-filen

Feilmeldinger skal skrives ut dersom filer mangler.

6. Opptelling av hvor mange elektroniske dokumenter (filer) som følger med

Her forutsettes det at hvert dokument avleveres som én fil. Opptellingen kan gruppere filene etter type: PDF, TIF eller TXT (ren tekst).

7. Kontroll av at filnavnene i referansene i NOARKIH-filen er identiske med navnene på filene i uttrekket

Alle filnavn skal skrives med store bokstaver. Bruk av små bokstaver skal gi feilmelding. De feilskrevne filnavnene skal listes opp. Dersom det refereres til tabellfiler som ikke finnes, skal de listes opp (jf. punkt 5).

Hvis filnavnene ikke stemmer overens med hva som er definert for Noark-4, vil det føre til feil senere i programmet ved kontroll av fremmednøkler.

Kravspesifikasjon (6)

8. Kontroll av at alle dataelementene (attributtene) som er definert i NOARKIH-filen, forekommer i tabellfilene

Dataelementene behøver ikke å forekomme i alle postene i tabelluttrekkene. Det er nok at de forekommer minst én gang. Dataelementer som ikke har noen forekomster, skal listes opp.

9. Kontroll av om det finnes dataelementer (attributter) i tabellfilene som ikke er definert i NOARKIH-filen

Dette innebærer en feil i uttrekket i forhold til Noark 4-standarden, og navnet på dataelementene skal listes opp.

Kravspesifikasjon (7)

10. Opptelling av antall poster i hver enkelt fil i tabelluttrekket, og kontroll av at dette stemmer med det antallet som er angitt i NOARKIH-filen

Antall poster i tabellfilene og antallene angitt i NOARKIH skal listes opp ved siden av hverandre.

11. Kontroll av at filreferansene i tabellfilen DOKVERSJON peker til dokumentfiler som er med i uttrekket

Filreferansen (VE.FILREF) skal inneholde hele stien til dokumentene slik de er plassert på avleveringsmediet, ikke slik de var plassert i det originale produksjonssystemet. Programmet har tatt høyde for at denne stien kan endres.

Dersom det refereres til filer som ikke finnes, skal disse telles opp og navnene skrives ut.

Kravspesifikasjon (8)

12. Kontroll av om det er med dokumentfiler i uttrekket som ikke har noen filreferanse i tabellfilen DOKVERSJON

Slike filer skal også telles opp og navnene skrives ut.

13. Opptelling av antall saker pr. år i tabellfilen NOARKSAK

Denne opptellingen skal grupperes etter årstallet i saksnummeret (SA.AAR).

14. Opptelling av antall journalposter pr. år i tabellfilen JOURNPOST.

Denne opptellingen skal grupperes etter årstallet i løpenummeret (JP.JAAR).

Kravspesifikasjon (9)

15. Opptelling av antall saker og journalposter (og eventuelt dokumenter) pr. år i rapportfilen SAKDOK

Denne opptellingen skal grupperes etter årstallene i henholdsvis saksnummeret

(SA. SAAR) og løpenummeret (JP.JAAR). Dersom elektroniske dokumenter også inngår i uttrekket, skal disse grupperes etter filreferansen (VE.FILREF).

16. Opptelling av antall journalposter pr. år i rapportfilen JOURNAL

Denne opptellingen skal grupperes etter årstallet i løpenummeret.

17. Analyse av hull i saksnr.rekkefølgen pr. år

Antall saker, første og siste saksnummer og antall hull skal listes opp. Alle hull større enn ti skal listes opp.

18. Kontroll av om det finnes saker utenfor den oppgitte arkivperioden

Start- og sluttdato for perioden angis før kjøring av programmet, og saksdatoen (SA.DATO) skal kontrolleres mot denne. Det skal ikke forekomme saker med dato etter sluttdatoen for perioden. Det kan forekomme saker med dato før startdatoen for perioden, dersom det ved forrige periodisering ble brukt mykt periodeskille.

<En tilsvarende kontroll av journaldatoen (JP.JDATO) i journalpostene kunne også ha vært utført. En slik kontroll utføres på Noark 3- uttrekk. Men denne kontrollen er ikke tatt med i denne kravspesifikasjonen.>

19. Import av alle tabellfilene i en relasjonsdatabase

Denne importen skal kjøres uten beskrankninger. Dersom det finnes rader som ikke lar seg importere, skal tabellnavnet og antall rader som ikke ble importert, listes opp.

20. Kontroll av at alle verdiene i primærnøkkelfeltene er unike

I noen tabeller i Noark-4 er det definert at en *kombinasjon* av flere felter (dataelementer) utgjør primærnøkkelen. Dersom det finnes nøkkelverdier som ikke er unike innenfor tabellen, skal navnet på tabellen og feltet/feltene listes opp.

Kravspesifikasjon (12)

21. Kontroll av at verdiene i fremmednøkkelfeltene samsvarer med verdier som forekommer i primærnøkkelfelter (eller andre kandidatnøkkelfelter) i de tabellene det refereres til

Dersom en fremmednøkkel ikke forekommer som primærnøkkel i den tabellen det refereres til, skal navnet på fremmednøkkelen og den tilhørende tabellen samt primærnøkkelen og den tilhørende tabellen listes opp. Antall forekomster (rader) med feil skal telles opp.

<På denne måten får vi bl. a. kontrollert at alle journalposter er tilknyttet reelle saker. Men vi får ikke kontrollert om alle saker har journalposter. Dette er en kontroll som utføres ved testing av Noark 3-uttrekk. Men denne kontrollen er ikke tatt med i denne kravspesifikasjonen.>

22. Kontroll av samsvar i poster og feltverdier mellom tabellfilene og rapportfilene SAKDOK og JOURNAL.

Det skal kontrolleres at alle saker i tabellfilene er med i SAKDOK, og omvendt. Det skal også kontrolleres at alle journalposter i tabellfilene er med i SAKDOK og JOURNAL, og omvendt. Samsvar mellom verdiene i noen utvalgte felt (bl.a. datoer) skal også kontrolleres. Funn av manglende samsvar skal listes opp.

<Dette er kontroller som utføres ved testing av Noark 3-uttrekk, og de er spesielt nyttige for å finne feil i uttrekksrutinene, siden tabellfilene og rapportfilene i uttrekket er basert på den samme databasen.>

Testrapport

Eksempel på 1. rapportside (standardisert) fra et reelt uttrekk:

ARKIVVERKET
RIKSARKIVET OG STATSARKIVENE

Testrapport for arkivuttrekk

Arkivskaper: Akershus universitetssykehus HF

Arkiv: Akershus universitetssykehus HF

Systemnavn: ePhorte versjon 2.0.4

Systemtype: Journal (Noark-4)

Arkivperiode: 01.01.2003 - 30.06.2004

Rapportdato: 23.03.2010

ARKIVVERKET
RIKSARKIVET

ArkN4-testerfaringer (1)

- **ArkN4 anvendte opprinnelig Dom-parser**

  Medførte lange kjøretider

  Tidsforbruk for ”små” kjøringer: Typisk 30 min.

  Tidsforbruk for ”middels” kjøringer (ca. 500 MB): Typisk 2,5 timer

  Tidsforbruk for ”store” kjøringer (over 500 MB)

  -Eks.: Sjøfartsdir. (3,7 GB) tok ca. 13,5 døgn, 12 døgn etter minneutvidelse (2 GB)

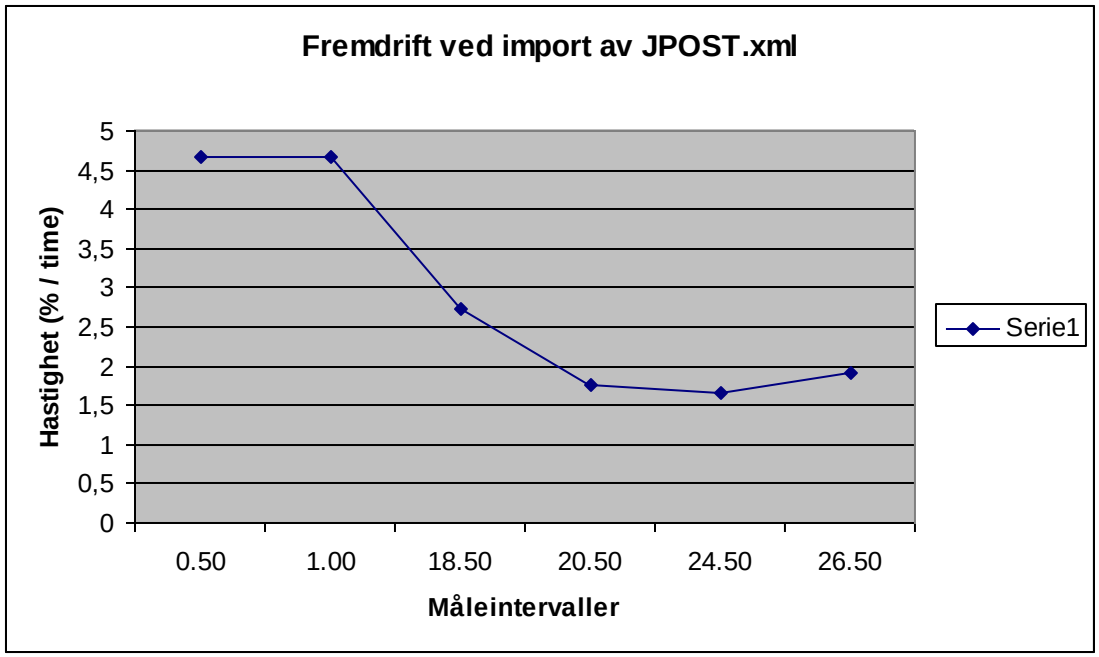
- **Tidsmålinger**

  - Foretok diverse målinger for å avdekke flaskehalser
  - Spesielt tidkrevende tabell: AVSMOT.XML
  - Etterfølgende eksempler viser fremdrift ved import av JPOST.XML (420 MB) og AVSMOT.XML (607 MB)

ArkN4-testerfaringer (2)

Fremdrift ved import av JPOST.XML

- Startmålepunkt relativt nær oppstart av import

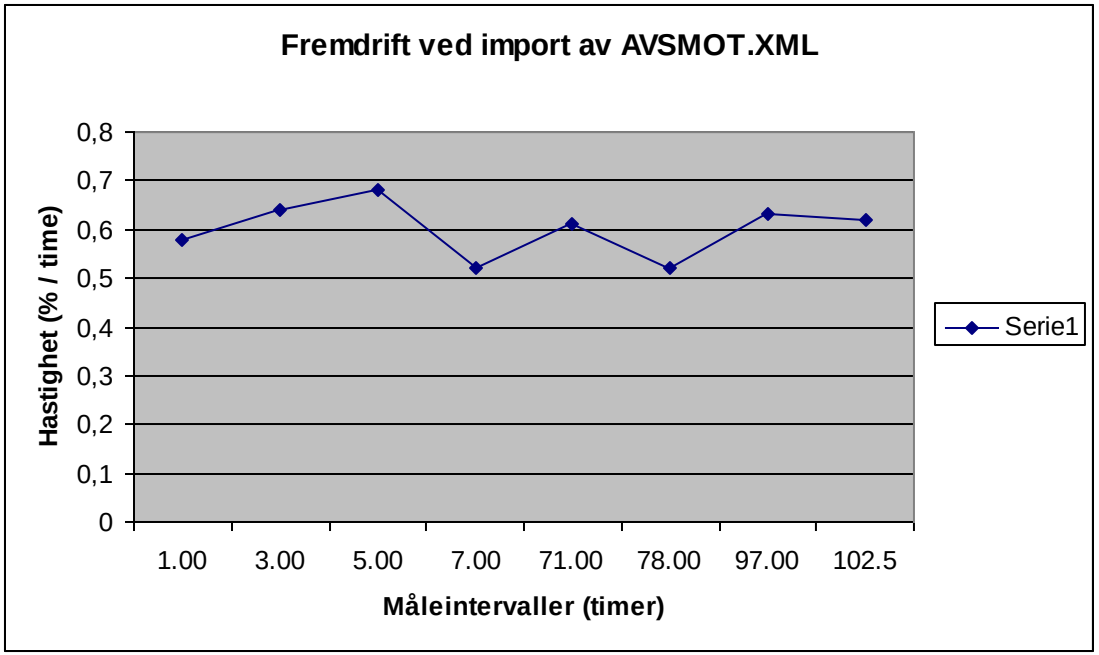


ARKIVVERKET
RIKSARKIVET

ArkN4-testerfaringer (3)

Fremdrift ved import av AVSMOT.XML

- Startmålepunkt mange timer etter oppstart av import



ArkN4-testerfaringer (4)

- **I ArkN4 versjon 1.3 ble skiftet parser til XML::Twig**
- **Tidsforbruk vesentlig redusert etter skiftet**

  Uttrekk på 1 GB går 7 ganger raskere

  Uttrekk på 3,7 GB går 28 ganger raskere
- **Tidsforbruk etter skiftet**

  Eks.:

  - Planleggings- og samordningsdep. (86 MB) 10 min.
  - Miljøverndep. (469 MB) 58 min.
  - Sjøfartsdirektoratet (3,7 GB) 10,5 timer

ArkN4-testerfaringer (5)

- **Kjørestopp**

  Årsaker til kjørestopp for Sjøfartdir. (AVSMOT.XML) grunnet parser:

  1 **Feilet p.g.a. X'1F' (US - Unit Separator) etter ca. 167 timer**

  2 **Skiftet til raskdisk/raskdisk: Feilet p.g.a. X'04' (EOT – End of Transmission) etter ca. 160 timer**

  3 **Feilet p.g.a. X'1F' (US - Unit Separator) etter ca. 168 timer**

- **En løsning:** Foreta egen test før ArkN4-kjøringen (RA har SAS-løsning)
- **Kan dette løses i parser?**

ArkN4-testerfaringer (6)

- **Parserbehandling av ”Whitespace” i XML**
  - **ArkN4 skriver følgende type feilmeldinger: ”x.x.1 Duplikat nøkkelverdi”**

    ‹OV.ORDNVER›233.2 VB‹/OV.ORDNVER› (mellomrom: 1 blank)
    ‹OV.ORDNVER›233.2 VB‹/OV.ORDNVER› (mellomrom: 2 blanke)
    ‹OV.ORDNVER›233.2VB‹/OV.ORDNVER› (mellomrom: 0 blanke)

  - **I nåværende versjon komprimerer parser TABS, SPACEBARS….. OK?**
  - **Kan ha konsekvenser for de 2 første variantene ovenfor**
  - **Den 3 varianten vises uendret av parser! Hva skjuler seg bak denne?**
- **Konklusjon: ArkN4 bør behandle alle tre variantene som forskjellige verdier og importere i databasen**
  - **Kan det settes ”flag” som styrer behandlingen i parser?**
  - **Andres erfaringer?**

ArkN4-testerfaringer (7)

- **Attributter med datatypen BOOL**
  - **ArkN4 skriver følgende feilmeldinger:**

```
6.7.1. Feil innhold i attributt
Attributtet 'AS.SPEFSAK' inneholdt en ugyldig verdi i forhold til datatypen 'BOOL'.
Filnavn              Linje    Beskrivelse
-----------------------------------------------------
ARSTATUS.XML    18      AS.SPEFSAK = -1(BOOL)
```

  - **Iht. Noark 4-kravspesifikasjonen skal felter som Arkn4 behandler som BOOLSKE ha verdi 0 eller 1**
  - **Valgt å behandle iht. Noark 4-kravspesifikasjonen: Avviser verdi –1!**

Utfordringer ved Noark 4-uttrekkstesting (1)

ArkN4 - Eksempler på faglige utfordringsområder:

Tegnhåndtering i XML- filer (ASCII- kontrollkarakterer avvises)

Parserbehandling av blanke tegn i nøkkelfelt

Testing av attributter med datatypen BOOL

Kontroll av tabeller kontra rapporter (kravspek. pkt. 22)

Kontroll av elektroniske dokumenter

Feil/forbedringsforslag beskrevet/dokumentert på TODO-liste (30 sider)

Noen feilårsaker

- Problematikk knyttet til konvertering før uttrekksgenerering (N-3 til N-4)
- Systemene har ikke implementert Noarks datamodell internt
- Forskjellige systemer har forskjellige metoder for å produsere utrekk

Utfordringer ved Noark 4-uttrekkstesting (2)

- **Testing av dokumentformater**
  - **Utføres ved hjelp av tilgjengelige verktøy?**
  - **Utvikle egne testverktøy?**
  - **De valgte verktøyene integreres med ArkN4?**

Kriteriene for å nekte godkjenning

Hva tester vi?

Hva er kriteriene for å nekte godkjenning?

§ 8-8 i Avlev.best. fastsetter kriteriene for å nekte godkjenning

Sentralt punkt: Feil eller mangler i dataene som kan tilbakeføres til feil ved selve produksjonen av uttrekket

Inkonsistens, feil og mangler i data må evt. være autentiske, dvs. de må også forekomme i det originale systemet.

Men det kan være tilfeller hvor vi må definere vårt ansvar snevert: til å bevare materialet slik vi mottok det

Inkonsistens og mangler – autentiske eller ikke – vil uansett kunne vanskeliggjøre vår senere fremstilling av bruksversjoner

Feilkilder (1)

I utgangspunktet er det 3 åpenbare feilkilder for feil i deponering/avlevering:

- Feil i datagrunnlag
- Feil ved uttrekksmetoden
- Feil i dokumentasjonen

Feilkilder (2)

Feil i datagrunnlag:

- 430 saker med arkivnøkler som ikke finnes i arkivnøkkelregisteret.
  ÅRSAK: Byttet arkivnøkler uten å periodisere
- Flere saker med samme saksnr.
- De kjente personene Donald Duck, Daffy Duck, osv.
- Obligatoriske felt mangler verdi.
- Innføring av nye journalposter etter at saken er avsluttet! Alt. tilbakeføring av avsluttet dato!

Feilkilder (3)

Feil i uttrekksmetode:

- I følgeskriv nevnes at antall journalposter skal være 166.097, mens det faktiske antallet er 42.285.
- Avleveringen inneholder saker som har datering utenfor angitt periode.
- Formatet på avleveringen følger ikke regelverket.
- Forskyvninger i forhold til gyldig format.

Feil i dokumentasjonen

Distribusjon av ArkN4

- **ArkN4 distribueres fritt!**
- http://www.arkivverket.no/ArkN4/
- Åpen kildekode-lisens
- GPL 2.0
  - the right to run the program, for any desired purpose
  - the right to study how the program works, and modify it. (Access to the source code is a precondition for this)
  - the right to redistribute copies
  - the right to improve the program, and release the improvements to the public
- Mål

  - få flere til å **teste** programmet.

  - teste arkivuttrekk før avlevering til Arkivverket